Proceedings Paper:
Filippou, V., Redmond, A. C., Bennion, J. et al. (2 more authors) (2020) Capturing accelerometer outputs in healthy volunteers under normal and simulated-pathological conditions using ML classifiers. In: 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society:Enabling Innovative Technologies for Global Healthcare, EMBC 2020. 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, EMBC 2020, 20-24 Jul 2020 Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS . IEEE , CAN , pp. 4604-4607.

https://doi.org/10.1109/EMBC44109.2020.9176201

Capturing accelerometer outputs in healthy volunteers under normal and simulated-pathological conditions using ML classifiers*

Filippou. V., Redmond. A.C., Bennion. J., Backhouse. M.R., and Wong. D.

Abstract— **Wearable devices offer a possible solution for acquiring objective measurements of physical activity. Most current algorithms are derived using data from healthy volunteers. It is unclear whether such algorithms are suitable in specific clinical scenarios, such as when an individual has altered gait. We hypothesized that algorithms trained on healthy population will result in less accurate results when tested in individuals with altered gait. We further hypothesized that algorithms trained on simulated-pathological gait would prove better at classifying abnormal activity.**

We studied healthy volunteers to assess whether activity classification accuracy differed for those with healthy and simulated-pathological conditions. Healthy participants (n=30) were recruited from the University of Leeds to perform nine pre-defined activities under healthy and simulated-pathological conditions. Activities were captured using a wrist-worn MOX accelerometer (Maastricht Instruments, NL). Data were analyzed based on the Activity-Recognition-Chain process. We trained a Neural-Network, Random-Forests, k-Nearest-Neighbors (k-NN), Support-Vector-Machines (SVM) and Naive Bayes models to classify activity. Algorithms were trained four times; once with 'healthy' data, and once with 'simulated-pathological data' for each of activity-type and activity-task classification.

In activity-type instances, the SVM provided the best results; the accuracy was 98.4% when the algorithm was trained and then tested with unseen data from the same group of healthy individuals. Accuracy dropped to 52.8% when tested on simulated-pathological data. When the model was retrained with simulated-pathological data, prediction accuracy for the corresponding test set was 96.7%. Algorithms developed on healthy data are less accurate for pathological conditions. When evaluating pathological conditions, classifier algorithms developed using data from a target sub-population can restore accuracy to above 95%.

Clinical Relevance— **This method remotely establishes health-related data of objective outcome measures of activities of daily living.**

I. Introduction

Physical activity (PA) significantly influences people's health and well-being, and helps prevent and delay onset of several chronic non-communicable diseases [1]. Several methods have been used previously to measure levels of activity in people. Such methods include large and expensive laboratory systems [2], and inexpensive, but time-consuming, subjective measures such as questionnaires, surveys and diaries [3].

Recent advances in commercial wearable technology has led to multiple devices that can enable PA to be assessed objectively. Of these, the accelerometer is commonly used for quantifying activity intensity and counting the number of steps [4]. Accelerometers are inexpensive, easy to use and long-lasting. However, common algorithms, including those used in consumer devices, are designed to be accurate for an archetypal healthy user and so may not be representative of subgroups such as those with chronic diseases that affect gait [5], [6]. Research to date has used accelerometers to classify activities and number of steps in moderately healthy patient populations [7], [8].

Our aim was to carry out a proof of concept study to investigate the performance of activity recognition algorithms using accelerometer data when trained on healthy individuals, but tested under healthy as well as unusual (*simulated-pathological*) gait conditions. We used a simulated-pathological condition, since recruiting actual patients was considered infeasible and impractical, especially given the exploratory nature of the current work.

We hypothesized that automated algorithms trained to identify types of physical activities in healthy participants would perform less well on participants when simulating a pathological gait.

II. Methods

A. Recruitment process

Participants were recruited via email and word of mouth from the staff and students of the University of Leeds. Participants were considered eligible for inclusion if they could walk freely without pain for two minutes. All participants were healthy, without any musculoskeletal condition or any condition affecting their gait. Participants 18+ years of age were recruited and, all participants gave informed written consent. Local ethical approval was provided by the University of Leeds (Ref #: MREC16-172).

B. Data acquisition

1) Data Sources

Each participant wore a MOX tri-axial accelerometer (Maastricht Instruments, Maastricht, NL) (dimensions:

*This study was supported by the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Tissue Engineering and Regenerative Medicine—EP/L014823/1.

Filippou. V is with the Institute of Medical and Biological Engineering, University of Leeds, Leeds, UK (phone: 07500481379; e-mail: mn12vf@leeds.ac.uk).

Redmond. A is with the Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK(email: A.Redmond@leeds.ac.uk).

Backhouse. M.R. is with the York Trials Unit, University of York, York, UK (email: mike.backhouse@york.ac.uk).

Wong. D is with the Centre for Health Informatics and Department of Computer Science, University of Manchester, Manchester, UK (email: david.wong@manchester.ac.uk).

Jacqueline. B is with the Royal Free London NHS Foundation Trust, London, UK (email: Jacqueline.Bennion@nhs.net)

35×35×10mm, weight: 11g). The device was held in place on the non-dominant wrist by an elasticated strap. The accelerometer had a measurement range of ±8g and a sampling frequency of 100 Hz. Recorded signals were stored locally on the accelerometer's internal memory (2GB) as a binary file that was downloaded upon the completion of each participant trial.

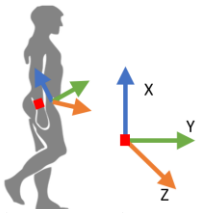

Figure 2: Accelerometer location and axis orientation

Our gold standard was a video recording of each participant. We used slow motion playback of videos to label the accelerometer data with the number of steps and to define the start and end time of each activity. This was cross-verified by an independent observer three times. The camera followed at approximately 2m from the participants.

2) Experimental protocol and set-up

Before attaching the activity monitor, participants were instructed that they would be performing nine activities: lie down, sit, stand, stand-to-sit, slow walk, normal walk, fast walk, walk upstairs, walk downstairs. Upon monitor attachment, the participant was asked to jump once to facilitate alignment of the video and accelerometer. After the jump, the participant performed the nine activities sequentially, and was reminded of each task. Participants were asked to jump once again after activities had been completed.

Each set of activities were performed twice, once under healthy conditions, and once under simulated-pathological conditions. For the simulated-pathological conditions, participants were asked to repeat the series of activities using a shuffling gait and to perform the activities more slowly. A shuffling gait was defined as when the foot is moving forward at the time of initial contact or during mid-swing, with the foot either flat or at heel strike, usually accompanied by shortened steps, reduced arm swing and forward flexed posture [9]. Such gait is a common marker of diseases such as severe rheumatoid arthritis and stroke. A written description, figure and video of shuffling gait was given to the participants prior to data collection. Participants were free to practice before data acquisition began.

C. Data processing

1) Data extraction

The binary files from the accelerometer were imported into Python™ (v3.6) for analysis. The extracted text files contained three columns of acceleration data, representing acceleration along the three principal axes.

To reduce the impact of high frequency random noise generated during data capture (caused, for instance, by muscle contraction), the accelerometer signal was filtered using a 6th order Butterworth filter with a 3Hz cutoff. The frequency of human activity is between 0-20 Hz and almost all of the signal energy is contained below 3 Hz [10]–[12].

We then derived five continuous signals from the 3-axis accelerometer data: dynamic accelerations, total magnitude, jerk, angular velocity and inclination angles.

Dynamic accelerations were calculated by averaging the readings on each direction, and then subtracting the corresponding average value from the raw acceleration signal. *Total magnitude* was calculated as:

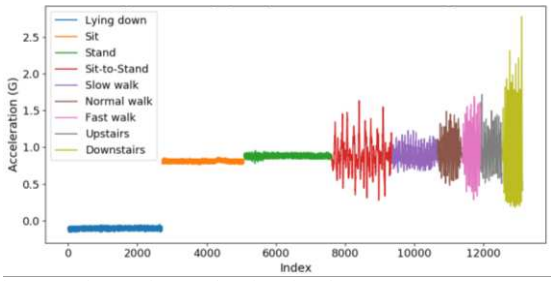

Figure 1: Time-series acceleration signal

$$acc = \sqrt{x^2 + y^2 + z^2}$$

Jerk is the rate of change of acceleration. A first order approximation was estimated from the acceleration signal as:

$$jerk = (acc_{t+T} - acc_t)/T$$

Where T is the sampling period. *Angular velocity* was identified by calculating the angle between the acceleration vectors in the current and the previous point. The accelerometer registers the data at equal time intervals. Therefore the angle between the vectors provides the angular velocity :

$$\cos(i, i+1) = \frac{(x_i x_{i+1} + y_i y_{i+1} + z_i z_{i+1})}{(\sqrt{x_i^2 + y_i^2 + z_i^2} \times \sqrt{x_{i+1}^2 + y_{i+1}^2 + z_{i+1}^2})}$$

Inclination angle was calculated for each direction.

$$\phi_x = \arccos(x^2/acc)$$

The continuous data were split into a series of short time windows, in which the signal may be approximated as stationary. We used windows of 200 samples, corresponding to a time period of 2 seconds, exceeding the Nyquist limit required to detect slower gait and within the range of window lengths proposed in prior research [13].

Each window was manually labelled with a specific activity task and assigned to one of three broader activity types (static, dynamic, transition) using the video gold standard. Each activity task corresponded to an activity type. Dynamic activity tasks were slow walk, normal walk, fast walk, ascending and descending stairs. Static activity tasks were lying, sitting, standing. The transition activity type comprised the stand-to-sit task only.

2) Feature extraction and selection

From the acceleration time series in each window, we extracted a set of 120 summary features to represent the acceleration (x, y, z, t), jerk (x, y, z, t), angular velocity and inclination angle (x, y, z) signals. The features derived were time-domain (mean, standard deviation, kurtosis, skewness, root mean square, interquartile range, power spectral density) and frequency-domain (energy, max frequency, max 2nd frequency, mean frequency, entropy). A reduced number of linear combinations of these features were selected using principal component analysis (PCA). A cut-off total explained variance of 0.95 was set on the explained variance. By reducing the dimensionality of the feature set, we limited the risk of overfitting subsequent classification models. The features were reduced to 25 and 30 principal components for healthy and *simulated-pathological* groups respectively.

Table 1: Machine learning (ML) algorithm parameters

Parameters	Activity type		Activity task	
	H/H	S/S	H/H	S/S
k-NN (K neighbors)	4	4	4	4
NN (neurons)	35	55	60	75
RF (trees, min_samples_split*)	4, 12	4, 12	4, 12	4, 12
SVM (C, gamma)	1, 1	10, 1	1, 1	10, 1
* minimum number of data required to split an internal node				

The PCA feature set was then used as input to a selection of five machine learning classifiers: back-propagation Neural Networks (NN), Random Forests (RF), Support Vector Machines (SVM), k-Nearest Neighbours (kNN), and Naive Bayes (GB). These classifiers have been commonly used for clinical classification problems [14]–[16]. The parameters used for each algorithm are shown in table 1.

Each classifier was assessed on its ability to classify both activity type and activity task. We conducted three algorithmic scenarios:

1. trained on healthy data; tested on healthy data (H/H)
2. trained on *simulated-pathological* data; tested on *simulated-pathological* data (S/S)
3. trained on healthy data; tested on *simulated-pathological* data (H/S).

Performance was assessed using accuracy [14]. For scenarios (1) and (2), performance was estimated using 10-fold cross validation, and we report the mean performance. For scenario (3) all relevant data were used for training and testing.

III. Results

The mean age of participants was 32.7 years (s.d 12.7). Of the 30 participants, 14 identified as female. Their mean height was 171.5 cm (s.d 7.1) and their mean weight was 69.2 kg (s.d 13.6).

The highest level of accuracy for activity classification was achieved using SVM and k-NN in activity-type and activity-task groups respectively (Table 2). All ML approaches demonstrated higher accuracies for the broader activity-type identification than for specific activity-task identification. The SVM and k-NN classifiers achieved an accuracy of 98.4% and 94.3% for activity-type and activity-task identification respectively in classifiers trained on healthy data (H/H). When these classifiers were applied to *simulated-pathological* data (H/S), to replicate real world use of wearable accelerometers, accuracy fell between 31.3%-52.8%. Training the algorithms using simulated pathological data and then identifying simulated pathological activities (S/S) improved the accuracy to 96.7% and 84.5% for activity-type and activity-task identification respectively.

Confusion matrices are performance measurements which were developed to visualize accuracy and other metrics (figures 3-4). Figure 3 shows that static, stand-to-sit and slow walk activities achieved high individual recall scores, with lying achieving the highest recall score as 0.996. Fast walk obtained the worst recall performance, which was 0.796. In terms of the precision score, static, stand-to-sit and downstairs activities achieved scores greater than 0.940. Normal walk obtained the worst precision score which was 0.798. Figure 4 demonstrates that static activities had the three greatest recall scores, while lying, sitting and stand-to-sit had the three highest precision scores.

Table 2: Machine learning algorithm evaluation (accuracy)

ML algorithms	Group (Train/Test)		
	H/H	S/S	H/S
	Activity type: Static, Dynamic, Transition		
NN	0.983(0.982-0.983)	0.957 (0.956-0.958)	
RF	0.953 (0.952-0.954)	0.921 (0.920-0.923)	
k-NN	0.983 (0.982-0.983)	0.960 (0.959-0.961)	
GB	0.897 (0.896-0.898)	0.834 (0.832-0.836)	
SVM	0.984 (0.983-0.984)	0.967 (0.966-0.968)	0.528
	Activity task: Specific activities		
NN	0.926 (0.924-0.927)	0.770 (0.767-0.772)	
RF	0.873 (0.871-0.875)	0.689 (0.687-0.691)	
k-NN	0.943 (0.941-0.944)	0.845 (0.843-0.846)	0.313
GB	0.749 (0.746-0.751)	0.516 (0.514-0.518)	
SVM	0.926 (0.925-0.928)	0.838 (0.836-0.840)	

IV. Discussion

Earlier studies have attempted activity recognition using machine learning classifiers similar to those used here. In healthy volunteers, results were similar. All classifiers that were tested, except Naïve Bayes, had accuracies ranging from 68% to 98% [7], [14], [17]–[21]. Naïve Bayes provided poorer results than the other algorithms [14], [17]–[20].

Our results demonstrated high levels of accuracy when the classifier was trained and tested with data from a similar group. However, when the tested data (*simulated-pathological*) differed from the training data (healthy), the accuracy dropped dramatically.

The difference in mean accuracy is likely due to the fact that volunteers were asked to make significant changes to their motions under *simulated-pathological* conditions. Although we attempted to train participants to replicate compromised motion, we could not be certain that their movements accurately reflected real pathological motion. Indeed, participants may have interpreted the instructions on how to mimic the pathological activities slightly differently. This means that the accuracies reported can only be considered a reasonable initial estimate of the performance of ML algorithms on real patients.

Previous studies have assessed whether algorithms trained on data from healthy populations were suitable for pathological populations. They conclude, like us, that large differences between groups means that algorithms should be trained for specific target groups [22]–[24].

One potential limitation is that we have reported accuracies as our overall performance metric. It is well known that accuracy can be a poor metric of overall performance in the presence of unbalanced data.

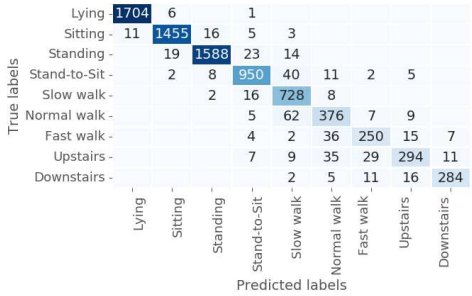

Figure 3: Confusion matrix of H/H group for tasks of activity

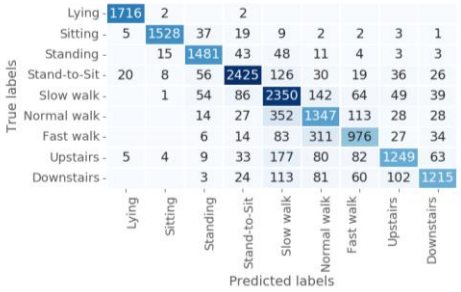

Figure 4: Confusion matrix of S/S group for tasks of activity

This problem was of lower concern here, in which quantity of each of activity was similar. Reporting the accuracy also allowed direct comparison to other work, and exact classifications are shown in table 2.

Another limitation in this study is human error for labeling the activities. Even though there was a gold standard video, the activity labeling was completed manually and subject to human error. However, the authors ensured thorough steps were taken to minimize this by using slow motion analysis, replaying analysis and triple counting each activity set.

One aspect of the study was to act as a baseline for developing activity classifiers for the healthy population. These classifiers will be further updated to suit the pathological population with walking impairments, and used by clinicians to evaluate the daily activity performance of chronic condition patients. Clinicians will be able to have a more informed view about the activity of their patients, hence provide them with better and patient-specific treatment plans and medications.

Additionally, a range of different accelerometer devices could be used, and their results compared to check the accuracy of the devices and the wider utility of the ML approach.

V. Conclusion

In this study, we used five machine learning algorithms to classify nine daily living activities. Activities were performed by healthy and *simulated-pathological* conditions. Furthermore, activities were classified into two groups, general activity type and specific activity task. The SVM and k-NN classifiers outperformed all other algorithms in activity-type and activity-task classifications respectively. All algorithms performed well when the training and test sets both came from the same population. Conversely, when the algorithms were trained with healthy data and tested with simulated-pathological data, as would usually occur in the real-world, the accuracy demonstrated was poor. It may therefore be possible to develop more accurate and clinically useful activity classification algorithms based on a person's or a sub-population's accelerometer gait signal.

References

[1] S. M. Phillips, L. Cadmus-Bertram, D. Rosenberg, M. P. Buman, and B. M. Lynch, "Wearable technology and physical activity in chronic disease: opportunities and challenges," *Am. J. Prev. Med.*, vol. 54, no. 1, pp. 144–150, 2017.

[2] S. J. Strath *et al.*, "Guide to the assessment of physical activity: Clinical and research applications," *Circulation*, vol. 128, no. 20, pp. 2259–2279, 2013.

[3] J. M. Broderick, J. Ryan, D. M. O. Donnell, and J. Hussey, "A guide to assessing physical activity using accelerometry in cancer patients," *Support Care Cancer*, vol. 22, no. 4, pp. 1121–1130, 2014.

[4] R. K. Walker, A. M. Hickey, and P. S. Freedson, "Advantages and limitations of wearable activity trackers: Considerations for patients and clinicians," *Clin. J. Oncol. Nurs.*, 2016.

[5] M. R. Backhouse, E. M. A. Hensor, D. White, A. Keenan, P. S. Helliwell, and A. C. Redmond, "Concurrent validation of activity monitors in patients with rheumatoid arthritis," *Clin. Biomech.*, vol. 28, no. 4, pp. 473–479, 2013.

[6] P. J. Mancuso, M. Thompson, M. Tietze, S. Kelk, and G. Roux, "Can patient use of daily activity monitors change nurse practitioner practice?," *J. Nurse Pract.*, vol. 10, no. 10, pp. 787–793, 2014.

[7] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist or ankle," *Med Sci Sport. Exerc*, vol. 45, no. 11, pp. 2193–2203, 2013.

[8] X. Kang, B. Huang, and G. Qi, "A novel walking detection and step counting algorithm using unconstrained smartphones," *Sensors*, vol. 18, no. 1, pp. 297–311, 2018.

[9] B. Salzman, "Gait and balance disorders in older adults," *Am. Fam. Physician*, vol. 82, no. 1, pp. 61–68, 2010.

[10] W. Z. Wang, B. Y. Huang, and L. Wang, "Analysis of filtering methods for 3D acceleration signals in body sensor network," *Int. Symp. Bioelectron. Bioinforma.*, pp. 263–266, 2011.

[11] M. Merryn, "Monitoring and interpreting human movement patterns using a triaxial accelerometer," 2003.

[12] E. K. Antonsson and R. W. Mann, "The frequency content of gait," *J. Biomech.*, vol. 18, no. 1, pp. 39–47, 1985.

[13] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.

[14] I. Cleland *et al.*, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors (Basel).*, vol. 13, no. 7, pp. 9183–9200, 2013.

[15] B. Erdaş, I. Atasoy, K. Açici, and H. Oğul, "Integrating features for accelerometer-based activity recognition," *Procedia Comput. Sci.*, vol. 58, pp. 522–527, 2016.

[16] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[17] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[18] Y. Saez, A. Baldominos, and P. Isasi, "A comparison study of classifier algorithms for cross-person physical activity recognition," *Sensors*, vol. 17, no. 1, pp. 66–91, 2017.

[19] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "How accurately can your wrist device recognize daily activities and detect falls?," *Sensors (Switzerland)*, vol. 16, no. 6, pp. 800–820, 2016.

[20] S. Zhang, A. V Rowlands, P. Murray, and T. L. Hurst, "Physical activity classification using the GENEA wrist-worn accelerometer," *Med. Sci. Sport. Exerc.*, vol. 44, no. 4, pp. 742–748, 2012.

[21] K. Lee and M. P. Kwan, "Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results," *Comput. Environ. Urban Syst.*, vol. 67, pp. 124–131, 2018.

[22] L. Lonini, A. Gupta, K. Kording, and A. Jayaraman, "Activity recognition in patients with lower limb impairments: Do we need training data from each patient?," *Eng. Med. Biol. Soc.*, pp. 3265–3268, 2016.

[23] N. A. Capela, E. D. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PLoS One*, vol. 10, no. 4, pp. 1–18, 2015.

[24] M. B. Del Rosario *et al.*, "A comparison of activity classification in younger and older cohorts using a smartphone," *Physiol. Meas.*, vol. 35, no. 11, pp. 2269–2286, 2014.